

Simple Search Engine Model: Adaptive Properties for Doubleton

Mahyuddin K. M. Nasution

Information Technology Department,
Fakultas Ilmu Komputer dan Teknologi Informasi
Universitas Sumatera Utara, Padang Bulan, Medan 20155, Sumatera Utara, Indonesia
mahyunst@yahoo.com, mahyuddin@usu.ac.id

Abstract. In this paper we study the relationship between query and search engine by exploring the adaptive properties for doubleton as a space of event based on a simple search engine. We employ set theory for defining doubleton and generate some properties.

Keywords: singleton space, search term, query, jaccard coefficient

1 Introduction

A search engine is extensively important to help users to find relevant information in Web. The search engines have different features, among of them are to service the tasks and subtasks that directly or indirectly uses the techniques such as indexing, filters, hub, page rank, hits, and etc [1], but to access any information in Web the users need search term and other literal text in a query. In this case, the query has become the leading paradigm to find the information, whereby the information retrieval (IR) is concerned with answering information need as accurately as possible. However, the difficult formulating of a query is always with the lack of understanding to special cases about the important information. The objective of this paper, therefore, to generate some adaptive properties of doubleton as semantic meaning of relation between a query and a search engine.

2 Related Works and Motivation

In literal text, a name means persons and personas (including pseudonyms), organizations, corporate, and government bodies and families, or any entity such as "Social Network" or like the literal text of "Superficial Method for Extracting Social Networks for Academics using Web Snippets" [2]. Any literal text or name, other case we called it as term, consists of words or tokens, a word w is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1, \dots, K\}$,

$$w_k = \begin{cases} 1 & \text{if } k \in K \\ 0 & \text{otherwise} \end{cases}$$

We defined some instances about a simple search engine [3].

Definition 1. A term t_x consists of at least one or a set of words in a pattern, or $t_k = (w_1 w_2 \dots w_l)$, $l \leq k$, k is a number of parameters representing word w , l is the number of tokens (vocabularies) in t_k , $|t_k| = k$ is size of t_k . ■

Definition 2. Let a set of web pages indexed by search engine be Ω , i.e., a set contains ordered pair of the terms t_{k_i} and the web pages ω_{k_j} , or (t_{k_i}, ω_{k_j}) , $i = 1, \dots, I$, $j = 1, \dots, J$. The relation table that consists of two columns t_k and ω_k is a representation of (t_{k_i}, ω_{k_j}) where $\Omega_k = \{(t_k, \omega_k)_{ij}\} \subset \Omega$ or $\Omega_k = \{\omega_{k_1}, \dots, \omega_{k_j}\}$. The cardinality of Ω is denoted by $|\Omega|$. ■

Definition 3. Let t_x is a search term, and $t_x \in \mathcal{S}$ where \mathcal{S} is a set of singleton search term of search engine. A vector space $\Omega_x \subseteq \Omega$ is a singleton search engine event (singleton space of event) of web pages that contain an occurrence of $t_x \in \omega_x$. The cardinality of Ω_x is denoted by $|\Omega_x|$. ■

Lemma 1. Let t_x and t_y are search term. If $t_x \neq t_y$, $t_x \cap t_y \neq \emptyset$ and $|t_y| < |t_x|$, then singleton search engine event of t_x and t_y is $\Omega_x = \Omega_x \cup \Omega_y$ or

$$|\Omega_x| = |\Omega_x| + |\Omega_y|, \quad (1)$$

where $\Omega_x, \Omega_y \subseteq \Omega$. ■

Lemma 2. If $t_y \neq t_z$ and $t_y \cap t_z = \emptyset$, then $|\Omega_y \cap \Omega_z| = 0$ and $|\Omega_y \cup \Omega_z| = |\Omega_y| + |\Omega_z|$. ■

Lemma 3. Let t_x and t_z are search terms. If $t_x \neq t_z$, $t_x \cap t_z = \emptyset$, and $\omega_x \cap \omega_z \neq \emptyset$, then $|\Omega_x| = |\Omega_z|$, $\Omega_x, \Omega_z \subseteq \Omega$. ■

One singleton space of event is not same to another if their search terms are not same. Two singleton spaces of event have a distance or a similarity. Let A be a set of search terms. A function $s : A \times A \rightarrow [0, 1]$ is called *similarity (proximity)* on A if s is non-negative, symmetric, and if $s(t_x, t_y) \leq s(t_x, t_x)$ holds for all $t_x, t_y \in A$, with equality if and only if $\Omega_x = \Omega_y$ [4,?].

In the context of modal logic, the similarity of two event spaces Ω_x and Ω_y for $\omega \Rightarrow t_x$ (true) and $\omega \Rightarrow t_y$ (true) respectively we use to explore the semantic relation of Ω where each search term is represented by a set of features [6]. Let two different search terms $t_x \neq t_y$ for representing a same entities, Ω_x be most similar to Ω_y where t_x is true, then $t_x \Rightarrow t_y$ will be true at Ω_y if and only if t_y is true at Ω_x , that is $\Omega_y(t_x) = 1$ if t_x is true at Ω_y then we have

$$\Omega_y(t_x \Rightarrow t_y) = \Omega_x(t_y) \quad (2)$$

where $\Omega_x(t_y) = 1$ if t_y is true at Ω_x . Similarly, by symmetry on a similarity, we obtain also

$$\Omega_x(t_y \Rightarrow t_x) = \Omega_y(t_x). \quad (3)$$

and to generate a similarity of two singleton spaces of event, the singletons associate with a doubleton space of event. We define a doubleton space of event as follows.

Definition 4. Let t_x and t_y are two different search term, $t_x \neq t_y$, $t_x, t_y \in \mathcal{S}$, where \mathcal{S} is a set of singleton search term of search engine. A doubleton search term is $\mathcal{D} = \{\{t_x, t_y\} : t_x, t_y \in \Sigma\}$ and its vector space denoted by $\Omega_x \cap \Omega_y$ is a double search engine event (doubleton space of event) of web pages that contain a co-occurrence of t_x and t_y such that $t_x, t_y \in \omega_x$ and $t_x, t_y \in \omega_y$, where $\Omega_x, \Omega_y, \Omega_x \cap \Omega_y \subseteq \Omega$. ■

For example, one of widely the used measures for generating the relations in a social network extraction is Jaccard coefficient [7], by using the singleton space and double space of events we have

$$s_{jc}(t_x, t_y) = \frac{|\Omega_x \cap \Omega_y|}{|\Omega_x| + |\Omega_y| - |\Omega_x \cap \Omega_y|} \quad (4)$$

Similar to singleton space of event for t_x and t_y in a query, we have

$$\begin{aligned} \Omega_x \cap \Omega_y &= (\Omega_x(t_x) = 1) \wedge (\Omega_y(t_y) = 1) \\ &= (\Omega_x(t_y) = 1) \wedge (\Omega_y(t_x) = 1) \\ &= (\Omega_x(t_x, t_y) = 1) = (\Omega_y(t_x, t_y) = 1) \\ &\supseteq \emptyset, \end{aligned} \quad (5)$$

thus

$$|\Omega_x \cap \Omega_y| = \sum_{\Omega} (\Omega_x(t_x, t_y) = 1) = \sum_{\Omega} (\Omega_y(t_x, t_y) = 1). \quad (6)$$

So we obtain

$$|\Omega_{x_p} \cap \Omega_{y_p}| = \sum_{\Omega} (\omega_{x,y} \Rightarrow t_x, t_y) \leq |\Omega_x \cap \Omega_y| \quad (7)$$

Therefore, the automatic extraction of meaning from the Web through using Ω affected by problem of singleton, i.e. a consequence in doubleton space of event.

Problem 1. Let t_x and t_y are two different search terms. If $t_x \in \Omega_x \cap \Omega_y$, then $t_x \in \Omega_x \cap \Omega_x$ and $t_x \in \Omega_y \cap \Omega_y$ such that

$$|\Omega_x \cap \Omega_y| \stackrel{?}{=} |\Omega_x \cap \Omega_y| + |\Omega_x \cap \Omega_x| + |\Omega_y \cap \Omega_y| \quad (8)$$

3 The Adaptive Properties of Doubleton in Search Engine

This Lemma 3 explains that problem of singleton be $|\Omega_x| = |\Omega_y|$ if and only if $t_x \neq t_y$ but $t_x, t_y \in \omega_x \wedge t_x, t_y \in \omega_y$. In other word, based on combining equations, $\Omega_x = \{(t_x, \omega_x)\} = \{(t_x, \omega_x \cup \omega_y)\} = \{(t_x, \omega_x) \cup (t_x, \omega_y)\} = \{(t_y, \omega_x) \cup (t_y, \omega_y)\} = \{(t_y, \omega_x \cup \omega_y)\} = \{(t_y, \omega_y)\} = \Omega_y$. This shows that the search terms may be different but they come from same web pages, and in this case they take the same meaning from web.

Based on Lemma 1, $|\Omega_x \cap \Omega_y| = |\{(t_x, \omega_x)\} \cap \{(t_y, \omega_y)\}| = |\{(t_x \cap t_y, \omega_x \cap \omega_y)\}| = |\{(t_y, \omega_y)\}| = |\Omega_y|$ or

$$|\Omega_x \cap \Omega_y| = |\Omega_y| \quad (9)$$

Because $|\Omega_y| < |\Omega_x|$, we have $|\Omega_x \cap \Omega_y| < |\Omega_x|$. However, by Lemma 2, $|\Omega_x \cap \Omega_y| = |\{(t_x, \omega_x)\} \cap \{(t_y, \omega_y)\}| = |\{(t_x \cap t_y, \omega_x \cap \omega_y)\}| = \emptyset$. This means that

$$|\Omega_x \cap \Omega_y| < |\Omega_x| \wedge |\Omega_x \cap \Omega_y| < |\Omega_y|. \quad (10)$$

Based on Lemma 3, $|\Omega_x \cap \Omega_y| = |\{(t_x, \omega_x)\} \cap \{(t_y, \omega_y)\}| = |\{t_x \cap t_y, \omega_x \cap \omega_y\}| = |\{(t_x, \omega_x)\}| = |\Omega_x|$ or

$$|\Omega_x \cap \Omega_y| = |\Omega_x| \quad (11)$$

Therefore, Eqs. (9), (10) and (11) clearly give $|\Omega_x \cap \Omega_y| \leq |\Omega_x| \leq |\Omega|$ or $|\Omega_x \cap \Omega_y| \leq |\Omega_y| \leq |\Omega|$, and this has proved the following theorem.

Theorem 1. *Let t_x and t_y are search terms. If $t_x \neq t_y$, but $\{(t_x, \omega_x)\} \cap \{(t_y, \omega_y)\} \neq \emptyset$, then a doubleton search engine event of t_x and t_y is the $\Omega_x \cap \Omega_y$, $\Omega_x, \Omega_y \subseteq \Omega$, $|\Omega_x \cap \Omega_y| \leq |\Omega_x| \leq |\Omega|$ and $|\Omega_x \cap \Omega_y| \leq |\Omega_y| \leq |\Omega|$. ■*

Disagree with it, let t_x and t_y are any search terms and we can derive a formula, that is, it starts from Eq. (11),

$$\begin{aligned} |\Omega_x \cap \Omega_y| &= |\Omega_x| \\ &= |\Omega_x| + |\Omega_y| && \text{Lemma 1} \\ &= |\Omega_x| + |\Omega_x \cap \Omega_y| && \text{Eq. (9)} \\ &= |\Omega_x| + |\Omega_y| + |\Omega_x \cap \Omega_y| && \text{Lemma 1} \end{aligned}$$

and we know that $|\Omega_x| = |\Omega_x \cap \Omega_x|$ and $|\Omega_y| = |\Omega_y \cap \Omega_y|$, then Eq. (8) in Problem 1 be

$$|\Omega_x \cap \Omega_y| = |\Omega_x \cap \Omega_y| + |\Omega_x \cap \Omega_x| + |\Omega_y \cap \Omega_y| \quad (12)$$

Thus Eq. (12) is a contraposition of Theorem 1. In other word, to reduce the enumeration of singleton and in order to Eq. (8) matches with Theorem 1, there should $t_y \in \Omega_y$ satisfies $t_y \neq t_x$, but $t_x, t_y \in \Omega_x$.

4 Conclusions and Future Work

Studying to properties of relation between query and search engine gave the understanding about the semantic representation of doubleton for object in literal text. Our near future work is to generate some selective properties of search engine.

References

1. W. B. Croft, D. Metzler, and T. Strohman. *Search Engines Information Retrieval in Practice*. Addison Wesley. 2010.
2. M. K. M. Nasution. Kolmogorov complexity: Clustering and similarity. *Bulletin of Mathematics*, 3(1): 1-16, 2011.
3. M. K. M. Nasution. Simple search engine model: Adaptive properties. arXiv:1212.3906, Cornell University Library: 2012.
4. M. K. M. Nasution and Shahrul Azman Noah. Superficial method for extracting social network for academics using web snippets. *Rough Set and Knowledge Technology*, LNCS - LNAI 6401, Springer-Verlag: 483-490, 2010.
5. M. K. M. Nasution and Shahrul Azman Noah. Extraction of academic social network from online database. In *IEEE Proceedings of 2011 International Conference on Semantic Technology and Information Retrieval* (STAIR'11).
6. M. K. M. Nasution and S. A. Noah. Information retrieval model: A social network extraction perspective. In *IEEE Proc. of CAMP 2012*: 322-326, 2012.
7. M. K. M. Nasution, S. A. Noah, and S. Saad. Social network extraction: Superficial method and information retrieval. In *Proceeding of International Conference on Informatics for Development* (ICID'11): c2-110-c2-115, 2011.